Introducing `wwm-german-18k` – Can LLMs Crack the Million? (Or Win at Least 500 Euros?)

Matthias Aßenmacher[1,2†], **Luis Karrlein**[1†], **Philipp Schiele**[3], **Christian Heumann**[1]

[1]Department of Statistics, LMU Munich, [2]Munich Center for Machine Learning (MCML),
[3]Stanford University, Department of Electrical Engineering

Correspondence: matthias@stat.uni-muenchen.de [†] Equal contribution

Abstract

Language-specific evaluation of large language models (LLMs) for multiple-choice question answering (MCQA) is an important means to test their abilities for a multitude of different dimensions. With a data set assembled from questions from the German variant of "Who Wants to Be a Millionaire?" we evaluate a set of German models and ChatGPT concerning factual/commonsense knowledge, syntactic abilities, and logical reasoning, amongst others. We contribute this new MCQA data set, extracted from the show's episodes and designed to evaluate the ability of models to answer this diverse range of questions. To ensure data quality, we describe our preprocessing, encompassing data cleaning, deduplication, and the creation of stratified splits. Furthermore, we fine-tune a set of German LLMs and prompt ChatGPT to provide baseline results. Our findings reveal that these models achieve (partly) satisfactory performance on questions of lower difficulty levels ($\leq$ 1000 euros). As the difficulty increases, performance steadily declines, highlighting the challenging nature of the later stages of the game. We contribute to the ongoing efforts to advance the capabilities of LLMs in comprehending and answering questions by providing a valuable resource for German MCQA research as well as further insights into the limitations of current LLMs.

1 Introduction

Recent advancements in transformer-based language models (Vaswani et al., 2017), especially with the advent of generative large language models (LLMs), such as OpenAI's GPT-series (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023), have demonstrated remarkable proficiency in various natural language generation and understanding tasks (Bubeck et al., 2023), including question answering (QA). LLMs are trained on vast amounts of text data from diverse sources, enabling them to learn language patterns, lexical semantics, and seemingly also factual knowledge. The exact extent to which e.g. factual knowledge is present in LLMs (and where exactly it is "stored" in the model weights) is still an open research question to be answered (Meng et al., 2022). As a result of the extensive pre-training, they exhibit impressive capabilities to apparently "comprehend" and respond to a broad spectrum of questions, making them potentially suitable candidates for tackling the challenging task of answering questions from quiz shows like "Wer wird Millionär?" (WWM; English: "Who Wants to Be a Millionaire?").

The QA task in the context of WWM represents an intriguing real-world use case for LLMs due to several compelling factors. First, this task requires not only the comprehension of questions but also the ability to reason, analyze answer choices, and make informed decisions. Second, when investigating the difficulty levels separately, we might gain more insights into how well LLMs can cope with different types of questions, either targeting solely factual knowledge or requiring more complex reasoning abilities. Moreover, the WWM format features questions across a wide range of domains, spanning from commonsense knowledge to more specific fields like science, sports, and pop culture. Consequently, an LLM capable of effectively answering such diverse questions must exhibit world knowledge, as well as factual accuracy, and must be able to grasp linguistic nuances across various topics. Thus, evaluating LLMs on this specific task can shed light on their knowledge representation capabilities and potential to handle multifaceted information.

Contributions: In this paper, we aim to investigate the feasibility and efficacy of employing German fine-tuned LLMs and ChatGPT for answering questions from the WWM quiz show. Our contribution is two-fold:

- We introduce a new multiple-choice question-answering (MCQA) resource for the German language allowing for a more comprehensive evaluation of German LLMs on this task. We (i) gather the data, (ii) extensively describe and motivate the pre-processing steps we applied, and (iii) provide a comprehensive descriptive analysis of the data.

- We evaluate the capabilities of different publicly available LLMs for this task and compare their performance across difficulty levels. This provides a reasonable baseline to compare against when evaluating ChatGPT on this task, even more so when generative LLMs with satisfactory capabilities for German emerge. Comparing fine-tuned to generative LLMs concerning their strengths and limitations in this context, we aim to contribute to the broader understanding of their capabilities and potential real-world applications in QA and game show formats.

2 Related Work: Other MCQA data sets for the German language

To the best of our knowledge, similar data sets from quiz shows or even "Who Wants to Be a Millionaire?" shows in other languages do not yet exist. When on the other hand considering language-specific related work and thus filtering the huggingface datasets space simultaneously for *German* and the task including "*multiple-choice-qa*", there are only eleven resulting data sets as of April 23, 2024.[1] All of these search results are, however, multilingual data sets and thus only a portion of the observations are in German. Other data sets in the German language in the realm of QA exist rather for the task of extractive QA[2], with `deepset/germanquad` (Möller et al., 2021) and `deepset/germandpr` (Möller et al., 2021) being probably the most prominent (purely German) examples. Nevertheless, none of these data sets is specifically aimed at evaluating German models *and* simultaneously targets MCQA. This stresses the need for a new data set for evaluating the ever-improving capabilities of modern-day LLMs.

[1] Search results as of April 23, 2024
[2] Search results as of April 23, 2024

3 The "wwm-german-18k" Data

3.1 Data Collection

The gathered data originates from the online version of the German quiz show "Wer wird Millionär?" (English: "Who Wants to Be a Millionaire?"), a format that is known across multiple languages. This iconic TV show, celebrated as one of Germany's premier and most recognized programs, challenges contestants with a series of fifteen questions. As they navigate through these questions, they stand a chance to win escalating monetary rewards, peaking at the coveted million Euro prize. These questions span a broad spectrum, from scientific inquiries to pop culture trivia, each of them accompanied by four potential answers and a constrained response time. As the quiz progresses, the complexity of the questions intensifies, but contestants are aided by specific lifelines, known as "Jokers", to facilitate their decision-making. To gather the data, we utilized web scraping techniques to engage with the online version of "Wer wird Millionär? Trainingslager"[3] (English: "WWTBAM? Training Camp"), hosted on RTL's website, the channel that airs the show in Germany. We initiated a game session by sending a POST request to the game's API. After establishing the session, we simulated individual games. For each game, we began at the first level corresponding to the 50 Euro prize. A random question for this level was then presented. Our script recorded the question along with its four possible answers. Importantly, regardless of the answer submitted, the system returned the correct one. This behavior aligns with the game's mechanics, where players are shown the right answer whether or not their guess was accurate. We then added this correct answer to our recorded data. Crucially, the game's training camp structure permits advancing to subsequent levels irrespective of the accuracy of the previous answer, ensuring a new question for each of the game's 15 levels can be drawn in each iteration. As we simulated numerous games, only new questions and their answers were added to our database. Given the assumption that questions are drawn independently, acquiring questions for each level mirrors the coupon collector's problem, where the goal is to collect all unique n items through m draws. We persisted in this iterative approach until reaching a point where new questions rarely emerged, sug-

[3] `https://spiele.rtl.de/spiele/rtl-spiele/wwm`

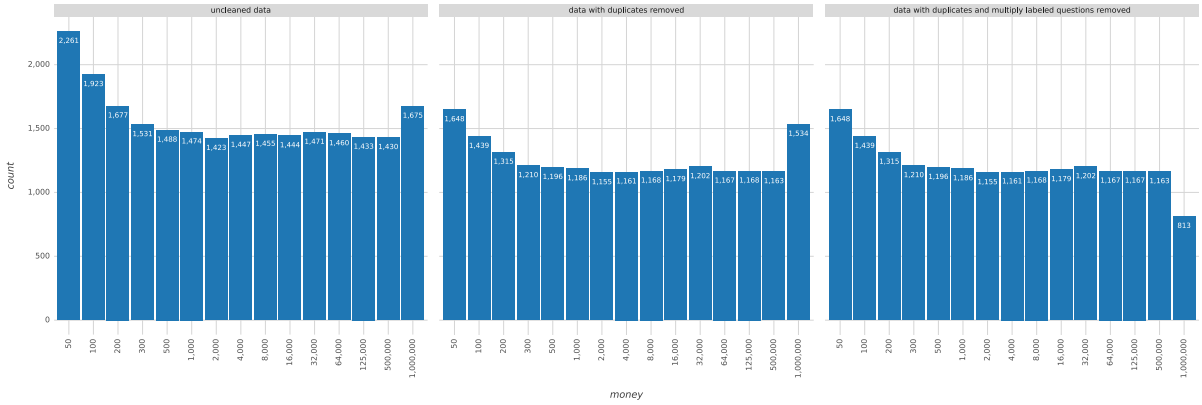

Figure 1: Comparison of the difficulty distribution in the different stages of processed data.

gesting we had captured the majority of available questions. We thus do not claim to have obtained an exhaustive collection of all questions, but rather a substantially representative collection of them that can be considered suitable for evaluating LLMs' capabilities.

3.2 Data Preparation

The unprocessed web-scraped dataset consists of 23,592 questions alongside all four possible answers with the right one as the "label" of the observation. Despite trying to avoid this during web scraping, there was a substantial amount of duplicates in the initial data, i.e. combinations of questions and answer options that occurred multiple times at different prize levels. For these data points, we applied deduplication and assigned the mean of the prize money categories to the new point, rounding down to a tiebreak; so if a data point occurred initially in the second and fifth category it would be put into the third category, and its duplicates removed.[4] Another difficulty encountered in the raw web-scraped data is that some questions came with multiple labels, i.e. multiple correct answer options. As this does not occur in the quiz which the data was taken from, where only one answer at a time is correct, these data points were deleted. These questions with multiple labels were disproportionally frequently present in the Million Euro questions which can be observed when comparing the three distributions in Figure 1. After discarding the erroneous data points, 18,169 deduplicated questions which we deem to be labeled correctly remain. As the final steps of the preprocessing, we sanitized the question endings as they included irrelevant (escape) characters, such as "\n" or spaces at the end of a question. We further unify occurrences of non-standard ellipses ("..", "...", "....") to a common form ("...") for the questions that have to be completed by the quiz show candidate. In three cases, we added missing question marks to the end of a question. The data also contained observations without spaces or ellipses at the end, however, these weren't grammatically complete sentences, but also required completion by one of the answer options. We thus keep them as they were.

Figure 2 illustrates the distribution of the endings, since the context is not always an actual question, but can also be an unfinished sentence that has to be completed. Ellipses ("...") or no ending ("") means the sentence is "cut off" and has to be completed by one of the options. Still, we observe that the majority of the "questions" are actual questions concluded by a question mark. The endings were extracted by using the following regular expression:

```
"(?<=[\w ÄaÖöÜüß%€$+-])[^\w ÄaÖöÜüß%€$+-]*$"
```

From descriptively analyzing the lengths[5] of the questions (i.e. the most important part of the model inputs) we learn that the distribution is notably skewed to the right (cf. Fig. 3). This is reflected by an arithmetic mean of the question length ($\bar{x} = 10.33$), exceeding the median question length ($\tilde{x}_{0.5} = 10$) by a margin of 0.33. While

[4] We acknowledge that this is somewhat heuristic, but after careful consideration, we think that this is an acceptable trade-off between the biases of either considering questions as too easy or too hard.

[5] We measure the length in *words*, since there is no uniform definition of tokens and lengths would thus differ from model to model.

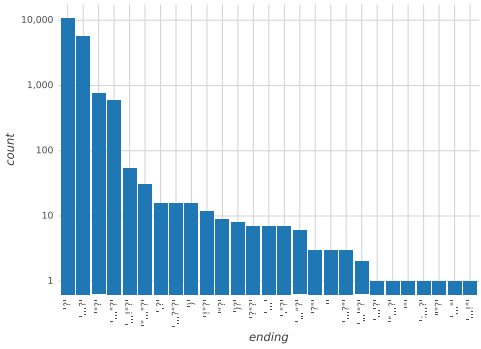

Figure 2: Absolute occurrence frequencies of the context endings in the data set.

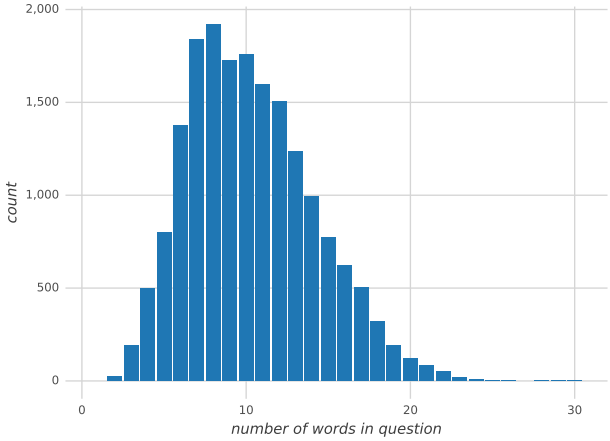

Figure 3: Distribution of the question lengths measured in words. Lengths measured in tokens may differ based on the specific model's tokenizer.

the majority of the questions exhibit a length of less than 10 words, less than 1% of them exceed a length of 20 words.

When subsequently analyzing the lengths of the different possible answers that are provided to the model alongside the question, we take on two perspectives: In Figure 4 we compare the lengths of all wrong answers to the respective correct answers. As there is virtually no difference between the two different distributions we can conclude that models will most likely not be able to pick up on such simple spurious patterns for solving the task. This observation is supported by the result of a Kolmogorov-Smirnov test exhibiting a p-value of $p = 0.95638$, leading to not rejecting the null hypothesis that the distributions do not differ.[6] Figure 4 reveals that the vast majority (approximately 59%) of the answers are only made up of a single word.[7] Nevertheless, there is also a substantial fraction of the answers consisting of two or more words (approximately 41%).

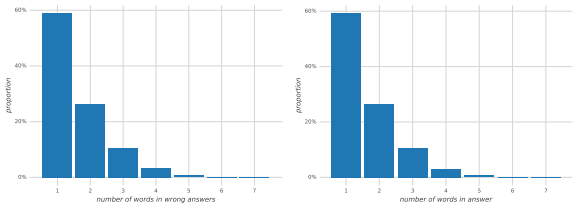

Figure 4: Lengths of wrong (left) and correct answers (right) aggregated across all difficulty levels.

The questions are split into 15 different strata, determined by the amount of money that can be achieved by answering them correctly (cf. Fig. 1 and Tab. 1). With increasing amounts of money to be won, the questions rise in difficulty: Starting from relatively simple ones, or most of the time even just simple puns, for winning 50€, the difficulty increases to ones that are presumably hardest to answer for achieving the 1,000,000€ prize. While for the simple (and often somewhat funny) puns for the first couple of difficulty levels frequently only one possible answer makes sense from e.g. a grammatical point of view, questions at higher levels in the quiz show tend to require either specific knowledge on niche topics or contain multiple plausibly-sounding answers to questions that sound relatively easy at first sight. One example of the latter is shown in Figure 5, where factual knowledge about wine is required. On the other hand, Figure 6 shows a question from the field of physics for a value of only 500 Euros. Here we observe that the answers may also relate to each other. In this case, they all have a connection to the concept of laziness or sluggishness (in German "*Trägheit*") which coincides with the German word for momentum (and is thus the correct answer). Another notable type of question occurs when the answer options themselves allow for excluding options (cf. Fig. 7). In this case three of the four options are distortions of words synonymous with offending or insulting a person, while only the second one, despite also *looking like* a distortion, is a really existing, though rather unknown, word with a different meaning.[8] In the quiz show, contestants try to answer one question from each prize money

[6] While this is no statistical proof that the distributions are equal, we at least were not able to show that the distributions differ significantly.

[7] Note, that this might result in more than one *token*, depending on the respective model's tokenizer.

[8] In the case of this particular option, this also certainly re-

money	question	label
50	Worauf radelt man zu zweit?	Tandem
100	Wer viel zu tief ins Glas geschaut hat, ist ...?	hackedicht
200	Teure Restaurants sind oft ...?	piekfein
300	Muss man löhnen, heißt es umgangssprachlich "Zahlemann und ..."?	Söhne
500	Ist es mit der Tugend nicht weit her, spricht der Volksmund von "Sodom und ..."?	Gomorrha
1,000	Was ist fein und glatt und neigt leider häufig zum Nachfetten?	Spaghettihaar
2,000	Wie wird in der Musikszene ein Auftritt bei einem Pop- oder Jazzkonzert genannt?	Gig
4,000	Welche Großstadt liegt nicht in Australien?	Auckland
8,000	Ist in den Medien vom Heppenheimer die Rede, dann geht es meistens um ...?	die Formel 1
16,000	Lindau am Bodensee liegt in ...?	Bayern
32,000	Wer soll gemeinhin mit sogenannten Genussscheinen angelockt werden?	Geldanleger
64,000	Was gewann Andreas Kofler gleich zu Beginn des Jahres 2010?	Vierschanzentournee
125,000	Was sieht aus wie ein Kolibri, ist aber ein Schmetterling?	Taubenschwänzchen
500,000	Bis 1958 hieß das Frisbee ...?	Pluto-Platte
1,000,000	Wo wurde Rainer Maria Rilke 1875 geboren?	Zürich

Table 1: Exemplary questions for each of the 15 difficulty levels.

class at a time, attempting to advance to the very last question to win the million euros. What is, however, of primary interest to us, is not whether an LLM could win the show, but how well it performs per prize money group. We thus perform a stratified train/validation/test split (80%/10%/10%) which ensures a similar balance of all 15 prize money categories across all three splits. For obtaining our baseline performances we utilize the validation set solely for early stopping during fine-tuning, while the test set allows for unbiased testing of the fine-tuned models.

4 Model Evaluation

4.1 Multiple-Choice Question-Answering

MCQA represents a pivotal challenge in natural language understanding and for the probing of factual knowledge. This task requires models to comprehend textual information from the given context or question and to select the most appropriate (i.e. the correct) answer option from a set of given choices, closely mimicking human reasoning and language comprehension abilities. MCQA plays a crucial role in many applications, ranging from exams in education and other assessment systems to information retrieval and chatbots. The complexity of this task lies not only in understanding the nuances of the question and the answer choices but also in grasping the context and potential ambiguities inherent in natural language. In quiz shows, these nuances and ambiguities can be ascribed to a pivotal role since questions or answer options are frequently intentionally created in a way that might confuse the contestant to some extent. Over the years, MCQA has evolved into a multifaceted research problem with applications to various domains (Hendrycks et al., 2020; Pal et al., 2022), incorporating various subtasks such as reading comprehension and semantic, logical (Srivastava et al., 2022), mathematical (Hendrycks et al., 2020), or commonsense (Srivastava et al., 2022) reasoning.

quires factual knowledge for humans. For well-trained LLMs, however, this could be easier as they might have seen the non-distorted word during pre-training.

4.2 Model architectures

In recent years, the field of MCQA has witnessed a remarkable transformation due to the advent of (generative) LLMs. There is a clear distinction between autoencoders, relying on discriminative fine-tuned task-specific modeling heads (such as BERT), and generative models that do not necessarily require fine-tuning (such as models from the GPT series). In our work, we rely on six German representatives of the former class of models, providing reasonable baseline values due to their proven and widely examined performance. Huggingface (Wolf et al., 2020) offers access to a wide range of pre-trained architectures via their model hub and allows for seamlessly integrating task-specific heads into the initial model architecture. For this analysis, we chose to use `bert-base-german-cased`, `bert-base-german-dbmdz-cased`, `bert-base-german-dbmdz-uncased`, `deepset/gbert-base`, `german-nlp-group/electra-base-german-uncased`, and `deepset/gelectra-base` alongside the `AutoModelForMultipleChoice` class from Huggingface. While BERT models (Devlin et al., 2019) represent the first large class of fine-tuned task-specific LLMs, ELECTRA (Clark et al., 2020) offers an alternative approach to pre-training, by

context:	Welche dieser Rebsorten ist Grundlage für renommierte Rotweine?	money:	32,000
options:	A: Cabernet Sauvignon, B: Chardonnay, C: Pinot grigio, D: Riesling	answer:	A

Figure 5: Question on the topic of winemaking

context:	Isaac Newton beschäftigte sich intensiv mit dem Prinzip der ...?	money:	500
options:	A: Müdigkeit, B: Bettruhe, C: Trägheit, D: Faulheit	answer:	C

Figure 6: Question from the field of physics

context:	Was macht eine Segelyacht, wenn sie sich zur Seite neigt?	money:	64,000
options:	A: peleidigen, B: krängen, C: spodden, D: ernietrigen	answer:	B

Figure 7: Question with non-words as options

focusing on token-level replacements. Pre-training BERT is mostly focused on the masked language modeling (MLM) task, where a percentage of 15% of the input tokens are corrupted and have to be subsequently predicted by the model. ELECTRA on the other hand employs the MLM task just as an intermediate step performed by an auxiliary generator model which creates predictions for the corrupted tokens and thus returns an ordinary text sequence. The actual ELECTRA model (the discriminator part of the training regime) takes the generator output as an input and is trained to predict for every token whether it is original or produced by the generator. Both models were initially proposed and trained for the English language, but relatively shortly after their release (purely) German versions for both architectures became available. We further examine the performance of ChatGPT (based on GPT-3.5 OpenAI, 2022) as one prominent representative of the class of generative LLMs.

4.3 Experimental Results

Our evaluation mostly focuses on providing reasonable baseline results for future research and differentiating model performance between the different difficulty levels among the questions. When comparing all of the six fine-tuned models and ChatGPT across difficulty levels (cf. Fig. 8 and 9), we observe the expected, relatively steady decline with increasing difficulty of the question (according to the prize money category) for BERT and ELECTRA (cf. Fig. 8), while ChatGPT exhibits constantly better performance for levels other than 1,000,000€ (cf. Fig. 9). Further, despite the overall performance decrease being rather consistent on average, there are still some irregularities. For some fine-tuned models, performance increases for one or two categories at some point on the difficulty scale, but without a clear pattern, and for the 300€ category there is a visible increase in performance *for all BERT/ELECTRA models* compared to the previous category. Overall it is important to keep in mind that an accuracy of 25% corresponds to random guessing, which is on average (nearly) the case for most of the higher prize money categories (also for ChatGPT). Concerning a comparison of the fine-tuned architectures, BERT vs. ELECTRA, models of the latter architecture (i) exhibit a higher average accuracy across all different difficulty levels which (ii) can be explained by better performance on especially the low-difficulty categories (below 2,000€). Model performance of ChatGPT proves to be very stable and high across most difficulty levels before it eventually starts to notably decrease at 64,000€ and exhibits a sharp drop for 1,000,000€. For the easier questions, there are only a few differences between the different fine-tuned models of the two underlying architectures, and we also do not observe notable differences to the performance of ChatGPT.[9] Nonetheless, we observe an overall performance difference between the two fine-tuned ELECTRA models. While deepset's `deepset/gelectra-base` achieves an accuracy of 53.83%, german-nlp-group's `german-nlp-group/electra-base-german-uncased` is better by a margin of 3.74% with a 57.57% accuracy. Another interesting observation is the decreasing variability in the accuracy with increasing question difficulty, thus decreasing overall model performance.

5 Discussion

Arguably, we do not (yet) use the data set to its full potential in this set of experiments, since we

[9] Nevertheless, one needs to keep in mind that ChatGPT is neither fine-tuned nor provided with few-shot examples.

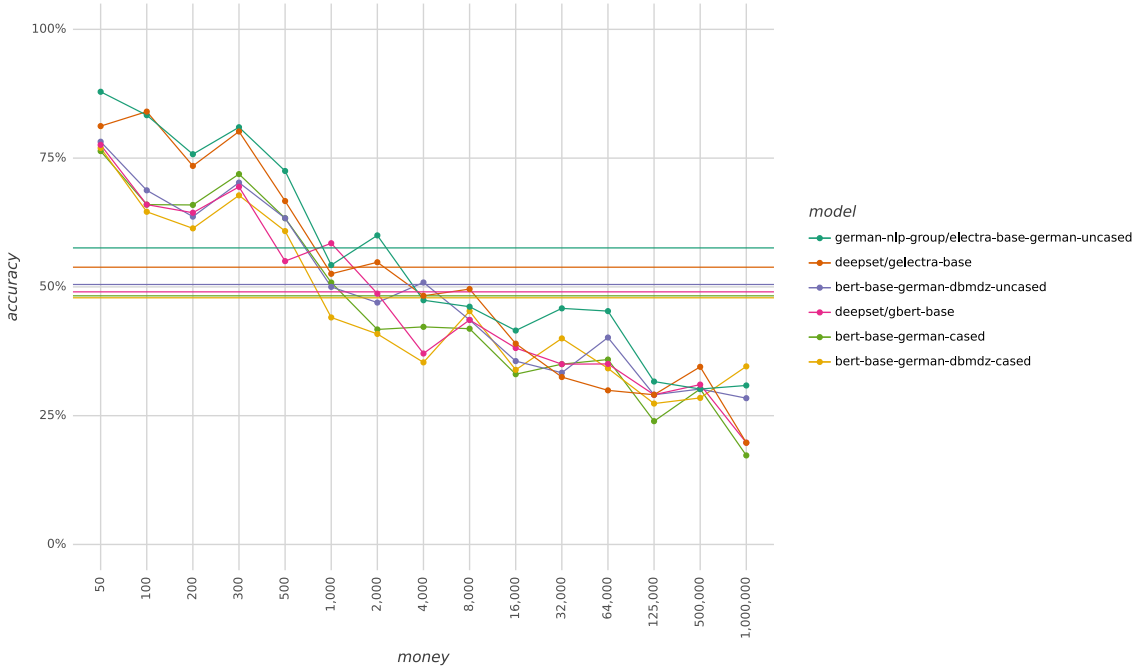

Figure 8: Accuracy of the evaluated models that were fine-tuned from different open-source models (separated by colors) split by the difficulty level of the questions (x-axis). Horizontal lines (in the respective colors) represent the model's average performance values across *all* difficulty levels.

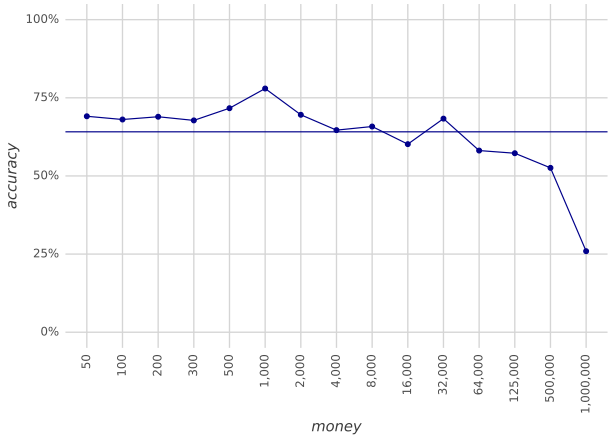

Figure 9: Accuracy of ChatGPT split by the difficulty level of the questions (x-axis).

only investigate the capabilities models can either *acquire* when fine-tuned on this MCQA data or the capabilities a generative LLM *already exhibits*. Beyond this use case, there's further potential for using the data in various few-shot learning settings, allowing for a more in-depth evaluation of prompting-based or generative LLMs. This few-shot setting would test the world knowledge and the reasoning capabilities already present in the LLMs' model weights, whilst nudging the model in the right direction, thus taking on another angle on this problem set. The results obtained from this evaluation of German BERT, ELECTRA, and ChatGPT on our newly introduced `wwm-german-18k` data set, however, still provide valuable insights into the strengths and limitations of heavily used LLMs in handling this large and diverse set of questions with varying levels of difficulty. The remainder of this discussion section will nevertheless shift the focus to potential enhancements that (open-source) generative LLMs can bring to solving MCQA tasks, along with a critical examination of the data set's potential shortcomings.

Recent developments in generative LLMs have led to remarkable performance when it comes to (seemingly) understanding and generating natural language text, which could also turn out to be a notable advantage for MCQA tasks. In response to closed-source models like ChatGPT, new generation of LLMs that is first and foremost characterized by openly-available weights emerged, preliminarily culminating in the publication of Llama3 (Meta, 2024) on April 18, 2024. Besides the "base" versions, many of these models are also released as quantized, instruction-tuned, or mixture-of-experts

versions allowing for (a) computationally cheaper adaption and (b) seamless usage of the models. Such models could potentially simultaneously benefit from the training example while exhibiting all the advantages that generative LLMs have over discriminative ones. This flexibility may empower them to also excel in tasks beyond MCQA, where answer choices are not explicitly provided, or when questions require generating more nuanced and contextually relevant responses. Additionally, generative LLMs could be leveraged for data augmentation purposes or for generating new, additional questions, thus enhancing the diversity and complexity of the data set. However, despite the diverse and interesting setting this data set is placed in, several potential shortcomings should also be acknowledged. The questions in the game show, and hence in the data set, may contain pop culture references, idiomatic expressions, or very specialized knowledge, which can pose challenges for both generative and discriminative models, especially when applied to more general domains. Additionally, the data set's focus on factual knowledge and trivia may not be fully adequate to evaluate the models' abilities in understanding and reasoning about more abstract or complex concepts holistically, which are arguably rather important for real-world applications.

Summing up, these debatable discussion points underscore the need for adequate resources to evaluate the promise of generative LLMs advancing the capabilities on MCQA, amongst others. The introduction of the `wwm-german-18k` data depicts an important step in that direction due to its challenging nature, for machine learning models and for humans. Simultaneously we also want to highlight the need for further data sets encompassing a broader range of question types and domains to further evaluate and refine these models. Future research needs to further aim at developing more diverse and contextually rich MCQA data sets that better represent the complexities of natural language understanding, ultimately driving the development of such data sets close to real-world scenarios will help to robustify LLM systems for MCQA across various languages and domains.

6 Conclusion

In conclusion, this research presents a dedicated and well-curated contribution to the field of German MCQA based on data extracted from the popular TV show "Wer wird Millionär?" alongside important baseline results for future research, showcasing one of the intended uses of the data: Evaluation of the progressing capabilities of LLMs. The primary contributions of this study can thus be summarized as follows:

First and foremost, we introduce a novel MCQA data set for the German language derived from the German version of the show "Who Wants to be a Millionaire?". This data set encompasses approximately 18,000 observations and thus provides a valuable resource for evaluating a diverse set of capabilities ascribed to modern-day LLMs. The diverse range of questions in the data spans various dimensions from factual/commonsense knowledge, over syntactic abilities to logical reasoning. To ensure the quality and reliability of our dataset, we carefully describe the careful preprocessing steps we took, which involved several aspects of cleaning the data, question deduplication, and the creation of stratified data splits. These steps are crucial for maintaining high data quality and providing a foundation for further research.

We also conducted extensive experiments using fine-tune two state-of-the-art German language models, namely German BERT and ELECTRA, as well as ChatGPT on our data set. The obtained baseline results offer insights into the performance of LLMs on this task, highlighting their competence in addressing questions with lower difficulty levels, up to approximately 1000€. However, as question complexity increases, our results reveal a consistent decrease in model performance, shedding light on the challenging nature of more difficult questions. This finding underscores the need for further research and model development to address these challenges and enhance MCQA performance on complex questions. Eventually, our contributions in the form of a new German MCQA dataset, detailed preprocessing methodology, and baseline results provide a valuable new resource for advancing the capabilities of German LLMs in comprehending and answering questions in natural language, particularly within the context of popular culture and entertainment. Further, it might serve as a blueprint for other languages, as this game show is popular around the world. This work invites researchers to build upon our findings and explore innovative approaches to improve the robustness and accuracy of MCQA models, ultimately contributing to the development of more robust and capable LLM-based systems.

Limitations

While we hope that this work provides researchers with a valuable non-English language resource for a more diverse evaluation of LLMs to gain more nuanced insights into their strengths and weaknesses, there are still issues we do not yet address in this work: First, we do not provide an exhaustive evaluation and comparison of different (open- vs. closed-source) generative LLMs, since this is not the focus of this work. Our focus is on the introduction of this new resource for comparing and evaluating LLMs. Second, this resource can also only be seen as a small contribution to the bigger question of how to properly benchmark generative LLMs. It only covers certain aspects of language and culture, but we hope this can serve as a valuable contribution to a better understanding of LLMs' behavior. Finally, as described in Section 3.2 there were some inconsistencies when recording the prize money category during web scraping, which we attribute to the subjectiveness of the concept of "difficulty" in the realm of quiz show questions. We would thus argue that our method for assigning the category can be regarded as a realistic approximation of the average perceived difficulty.

Ethical Considerations

We affirm that our research adheres to the ACL Ethics Policy. This work involves the use of data that is publicly available on the internet and does not involve harmful content or any personally identifiable information on humans. We declare that we have no conflicts of interest that could potentially influence the outcomes, interpretations, or conclusions of this research. All funding sources supporting this study are acknowledged in the acknowledgments section. We have made our best effort to document our methodology, experiments, and results accurately and are committed to sharing our code, data, and other relevant resources to foster reproducibility and further advancements in research.

Acknowledgements

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581.

Availability of the Data

Unfortunately, we are not able to share the data due to legal considerations: While scraping the data can, to the best of our knowledge, be considered legally fine, making it available to the public might lead to violations of copyright. We made several attempts to contact the production company EndemolShine Germany, explaining our endeavor to cooperate with them for publishing the data for research purposes only, but they were not open to any sort of conversation about this. Our legal council subsequently advised us against publishing the data without the company's explicit consent. Interested researchers are, however, encouraged to contact us for obtaining access to the data set.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv preprint arXiv:2104.12741*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.